5. Лексико-терминологические материалы для чтения текстов по психологии на английском языке: Частотный минимум / Сост. Г.В.Басовская, А.В.Вербицкий. Л., 1980.

6. Учебные терминологические материалы для чтения текстов по математике на английском языке: Частотный минимум/ Сост. Л.М.Сутягина. Л., 1982.

7. Частотный англо-русский словарь-минимум по квантовым генераторам / Сост. Н.С.Манасян. М., 1983.

8. Частотный англо-русский словарь-минимум по электронике / Сост. П.М.Алексеев. М., 1971.

Л.М.Сутягина
Уральский университет

ФОРМИРОВАНИЕ ВЫБОРОЧНОГО КОРПУСА ТЕКСТОВ ПРИ СОСТАВЛЕНИИ ЧАСТОТНОГО СЛОВАРЯ

(возможный алгоритм построения выборки)

В статье "Формирование выборочного корпуса текстов при составлении частотного словаря (качественная сторона формирования выборки)" [I], где обсуждались некоторые проблемы, связанные с качественной стороной формирования выборки для составления частотного словаря (ЧС), были сделаны следующие основные выводы:

I. Один из самых простых способов приблизить структуру выборки к структуре генеральной совокупности - это ограничить исследование как можно более узким подъязыком, т.е. выборка должна быть направленной;

2. Трудоемкую процедуру отбора текстов для составления ЧС с помощью жеребьевки или других аналогичных методов можно заменить направленным отбором ввиду практически случайного распределения лексических единиц (ЛЕ) в самих текстах;

3. К формированию узконаправленной выборки для составления ЧС целесообразно приступать лишь после того, как выделено "ядро" журналов и авторов в данной области знаний и определен по специальным таблицам полупериод жизни публикаций в этой области.

Достаточно ли выполнения этих условий для того, чтобы приступить непосредственно к составлению ЧС? Очевидно, нет. Более или менее ясно, какие тексты нужно включить в выборку, но не ясно, сколько нужно обследовать их для получения надежного

словаря, другими словами, неизвестен объем выборки.

Считается почти неоспоримым, что чем больше объем выборки, тем лучше. Однако возможности составителей небезграничны, и это приводит к тому, что на деле чаще всего либо пользуются известной формулой Р.М.Фрумкиной [2] (при этом создается видимость некоего статистического обоснования), либо останавливаются на ставшей уже практически традиционной цифре в 200 тыс. словоупотреблений, либо ограничиваются выборкой такой длины, обследование которой оказывается по силам. Все перечисленные способы определения объема выборки объединяет то, что они никак не учитывают ни широты тематики обследуемых текстов, ни особенностей конкретных подъязыков и стилей.

Можно было бы возразить, что существуют широко используемые методы проверки надежности данных ЧС и определения его эффективности[I], при помощи которых составители обычно подтверждают хорошее качество готового словаря, а следовательно, и достаточность объема выборки. Не останавливаясь на недостатках этих методов (критику существующего определения надежности данных ЧС см. напр., в [3, 4]), хотелось бы отметить, что проверяется, как правило, качество уже окончательно составленного словаря и никоим образом не прослеживается его связь с величиной объема обработанной выборки. Остается неизвестным, будет ли ЧС менее эффективен при меньшем объеме выборки или наоборот станет более действен, если объем выборки на много увеличить.

Проиллюстрируем сказанное на примере алгебры ЧС конечных групп (выборка в 100 тыс. словоупотреблений формировалась направленным отбором преимущественно из текстов с упоминанием о

[I] Под надежностью ЧС обыкновенно подразумевается отношение между относительной частотой ЛЕ в словаре и ее вероятностью в генеральной совокупности. По установившейся практике с этой целью рассчитываются доверительные интервалы для вероятностей ЛЕ при произвольно заданном уровне значимости, а также относительные ошибки определения границ этих интервалов. Эффективность обычно оценивается эмпирически - степенью покрываемости словарем аналогичного случайного текста, не вошедшего в выборку.

силовских p-подгруппах). Проследим, как изменяется эффективность частотного списка с увеличением объема выборки. Из приведенных в таблице данных видно, что, начиная с двадцатого текста, процент покрываемости каждого из вновь обследуемых текстов уже накопленным словарем не падает ниже 91. По-видимому, можно ожидать, что почти любой взятый наугад текст по алгебре конечных групп, имеющих силовские подгруппы, будет покрываться нашим словарем не менее, чем на 90%.

Таблица

Зависимость покрываемости каждого пятого по порядку рассмотрения текста выборки уже накопленным к моменту его анализа словарем от объема выборки

№ текста, п/п	Объем выборки, тыс. с/у	Покрываемость, %	№ текста, п/п	Объем выборки, тыс. с/у	Покрываемость, %
I	2	3	I	2	3
5	5	73	55	55	95
I0	I0	83	60	60	96
I5	I5	85	65	65	97
20	20	98	70	70	99
25	25	93	75	75	96
30	30	98	80	80	9I
35	35	95	85	85	99
40	40	95	90	90	97
45	45	9I	95	95	98
50	50	95	I00	I00	95

При этом следует отметить, что практически такого же результата мы достигли бы, ограничившись выборкой меньшего объема, т.е. для выделения лексического ядра выборки (самой частотной и равномерно распространенной лексики) такого узкого и лексически бедного подъязыка, как избранный нами, объем выборки в I00 тыс. словоупотреблений оказался избыточным. Естественно ожидать, что чем шире тематика, чем лексически богаче подъязык, тем большая длина выборки потребуется для достижения той же цели.

Тогда возникает вопрос: можно ли вообще с необходимой точностью оценить лексическое богатство подъязыка a priori

до соответствующей статистической обработки текстов, т.е. в конечном итоге, до составления ЧС? Такая возможность представляется едва ли осуществимой, и возникает замкнутый круг, разорвать который попытаемся, предложив такую методику составления ЧС, которая позволила бы контролировать достаточность проанализированного объема выборки непосредственно по ходу обработки текстов. Заметим, что это имеет смысл лишь в том случае, когда содержание ни одного из обследуемых текстов не выходит за рамки четко очерченной тематики, поскольку ясно, что всякое отклонение от заданной тематики влечет за собой регистрацию определенного количества "посторонних" для нас ЛЕ. Другими словами, можно сказать, что для этой цели выборка должна обладать межтекстовой лексической однородностью[2]. Остановимся на этом понятии более подробно.

Одной из черт, характеризующих текст и, соответственно, словарь является образование ядра в замкнутых группах, с одной стороны, и существование периферийных явлений - результат размытости системы, с другой [5; 6]. Ядро - это концентрированное выражение специфической лексики подъязыка. Концентрация лексики в начале частотного списка говорит, по-видимому, о том, что именно начало списка содержит большинство ядерных единиц, и чем более эта концентрация выражена, тем с более замкнутой системой (подъязыком) мы имеем дело. Периферийными явлениями в нашем случае являются редкоупотребительные (в пределе уникальные, т.е. встретившиеся в одном тексте выборки) ЛЕ.

Если как по количеству ядерных, так и по количеству уникальных ЛЕ тексты оказываются случайными, мы можем говорить об однородности наблюдений, о том, что обследованные тексты принадлежат одной и той же генеральной совокупности, а их лексика - лексике одного и того же подъязыка. Выборка в этом случае будет обладать межтекстовой лексической однородностью.

---

[2] Речь идет не о требовании полной однородности; полная однородность недостижима даже в очень тематически однородной выборке уже по тому, что кроме темы и ситуации вероятность выбора определенных слов в тексте обусловливается языковой компетенцией авторов [6].

Однако конечной целью анализа являются не первичные единицы-тексты, но вторичные - ЛЕ, содержащиеся в этих текстах. Требовать абсолютной однородности словаря, единицы которого безусловно являются разными, естественно, не имеет смысла, но можно, по-видимому, говорить об однородности словаря относительно признака принадлежности ЛЕ лексике исследуемого подъязыка.

Наличие межтекстовой лексической однородности выборки свидетельствует о том, что лексика текстов выборки, в целом, принадлежит одному и тому же подъязыку. Отсюда представляется возможным отождествление понятий лексической однородности выборки (и словаря) с понятием межтекстовой лексической однородности выборки.

Очевидно, чем у́же тематика выборки, тем больше ядерных и, соответственно, меньше неядерных ЛЕ содержит каждый ее текст, тем более случаен характер распределения последних. Это, в свою очередь, позволяет ожидать статистической устойчивости[3] числа уникальных единиц в тематически (а отсюда, и лексически) однородной выборке. Поэтому представляется возможным судить о степени лексической однородности выборки по степени статистической устойчивости числа уникальных ЛЕ в ее текстах.

Ядерные ЛЕ входят в большое число текстов выборки, стало быть для того, чтобы их зафиксировать, нужно просмотреть довольно малое количество текстов. Уже в первых по порядку рассмотрения текстах эти единицы будут встречаться. С этой точки зрения можно, очевидно, говорить о высокой скорости выделения ядерной лексики.

С другой стороны, вхождение ядерных ЛЕ в большое количество текстов (и как правило, с частотой большей, чем I) обеспечивает их попадание в разряд высокочастотных единиц. Естественно предполагать, что они заполнят верхние зоны частотного списка. Представляется маловероятным, чтобы неядерные элементы в тематически однородной выборке имели бо́льшую частоту, чем ядерные. Отсюда можно ожидать, что в ЧС ядерные ЛЕ будут располагаться более или менее подряд от начала списка, в основном не прерываясь неядерными. Это свойство ядерных ЛЕ назовем свойством четкой очерченности лексического ядра выборки (словаря).

---

[3] Статистическая устойчивость понимается в смысле Ю.И.Алимова [7].

Принимая во внимание сказанное, введем рабочее определение лексической однородности выборки [4].

Выборку будем считать лексически однородной, если

а) в ней найдется такая последовательность текстов, на которой ее лексическое ядро быстро выделяется;

б) ее лексическое ядро относительно четко очерчено;

в) ее периферийные явления-уникальные ЛЕ - обладают статистической устойчивостью.

В соответствии с данным определением возможные критерии лексической однородности выборки будут распадаться на две группы.

I. Критерии, относящиеся к выделению лексического ядра, - это критерии:

а) выявляющие четкость или нечеткость его очерченности;

б) выявляющие скорость его выделения;

2. Критерии, определяющие статистическую устойчивость периферийных явлений.

Для проверки гипотезы о лексической однородности выборки такие критерии были применены к данным уже упоминавшегося ЧС алгебры конечных групп. Произведенная с их помощью проверка показала наличие высокой степени лексической однородности этой выборки [3]. Не останавливаясь на деталях, скажем лишь, что лексическое ядро составило приблизительно 500 ЛЕ, которые концентрируются в начале частотного списка; они были в основном выделены на первых I5 тыс. словоупотреблениях (первых I5 текстах). Уникальные ЛЕ обнаружили хорошую устойчивость.

По-видимому, если по характеру стоящих перед ЧС задач необходимо, чтобы его выборка состояла из нескольких тематических разделов, то длины подвыборок, соответствующих каждому из разделов, должны быть достаточными для выделения лексических ядер этих разделов. Поскольку разделы выборок отраслевого ЧС обычно близки по тематике, то и пересечение лексических ядер его подвыборок будет довольно большим. Можно предположить, что под-

---

[4] Вводимое определение не претендует на универсальный характер, и предлагаемые ниже критерии ее проверки, очевидно, не являются единственно возможными.

выборки, проанализированные первыми, в какой-то мере "работают" на выявление достоверной части лексики последующих и поэтому могут иметь меньший объем, чем первоначальные.

Исходя из вышеизложенных соображений, особо важными в вопросе формирования выборочной совокупности текстов для составления ЧС представляются следующие моменты:

1) выборка должна быть представлена лексически однородными близкими по тематике разделами;

2) протяженность каждого такого раздела должна быть достаточной, но не излишне избыточной для выделения его лексического ядра;

3) количество разделов (примерный перечень их установлен заранее) при фиксированном объеме выборки определяется, видимо, непосредственно в ходе работы.

Ясно, что для того, чтобы указанные условия были осуществимы на практике, необходимо научиться устанавливать прямо в ходе работы над составлением ЧС, выделено уже лексическое ядро или нет. Большую помощь при оценке степени лексической однородности выборки и размеров ее лексического ядра оказывают критерии, описанные выше. Однако их использование требует предварительного обследования больших массивов текста, и поэтому они не могут быть непосредственно применены для решения стоящей перед нами задачи.

На основе данных о величине и темпе выделения лексического ядра нашей тематически и лексически однородной выборки был найден такой параметр, который позволяет достаточно оперативно судить о степени лексической однородности анализируемых выборок и, в частности, о темпе выделения лексического ядра.

Обозначим через k^* величину отношения количества новых (то же самое, что всех разных) слов в первом тексте к числу новых слов во втором, третьем и т.д. Последовательность таких отношений[5] $k^*_2, k^*_3, \ldots k^*_i, \ldots k^*_n$ должна по мере выделения лексического ядра выборки возрастать, а с момента его более или менее полного выделения ни одно из значений k^* не должно быть меньше некоторого порогового. Анализ последовательностей $\{k^*_i\}_{i=2}^{n}$, построенных на имеющемся в нашем распоряжении материале (а это, кроме

---

[5] Для удобства (чтобы i совпадало с соответствующим номером текста) нумерацию k^*_i начинаем с $i=2$

упоминавшегося ЧС, ЧС английского подъязыка электроники [9] и ЧС английского подъязыка топологии [8], позволил сделать вывод о том, что в качестве такого критического значения можно взять $k^*_{крит.} = 10$.

Подводя итог сказанному, можно предложить следующий алгоритм формирования выборки и непосредственного составления частотного словаря.

I. По имеющимся в наличии справочным источникам, типа УДК, ВИНИТИ, соответствующих реферативных журналов и т.д., а также в ходе консультаций со специалистами выяснить как можно более подробный перечень областей семантического пространства подъязыка, избранного для обследования; выделить наиболее значимый район выборки;

2. Исходя из физических возможностей, наметить примерный объем выборки. Практика составления частотных словарей подсказывает, что объемом, удобным для обработки одним исследователем, является 200 тыс. словоупотреблений. Кроме того, такой объем выборки стал своего рода стандартом для словарей группы "Статистика речи", что делает возможным сравнение ЧС и их последующее объединение;

3. Подобрать конкретные тексты наиболее значимого района выборки;

4. Установить единицу регистрации (словоформа, лексема, словосочетание);

5. На нескольких минимальных выборках заданного объема (обычно I тыс. словоупотреблений) установить среднее число регистрируемых единиц в одной минимальной выборке;

6. Начать последовательное расписывание текстов, при этом в каждом из них фиксировать число единиц, появляющихся впервые; одновременно вычислять значения k^*_i

$$k^*_i = \frac{m_{ср}}{m_i},$$

где $m_{ср}$ – среднее количество регистрируемых единиц в I тыс. словоупотреблений; m_i – число новых единиц в i – й тысяче словоупотреблений;

7. Обследование текстов, представляющих тематику данной области, прекращать, как только значения k^*_i станут устойчиво выше "десяти";

8. Руководствуясь составленным ранее (пункт I) перечнем, подобрать конкретные тексты по тематике смежного района выборки;

9. Продолжать обследование текстов по той же методике. Значение k_i^*, вследствие смены тематики, неизбежно вначале упадет, однако, благодаря тому, что пересечение лексических ядер двух разделов значительно, вновь возрастет до критического гораздо быстрее, чем при обследовании первой серии текстов;

I0. Последовательно переходить к новым разделам выборки, всякий раз выбирая наиболее близкий по тематике к предыдущему и прекращая расписывание, как только $k_i^* \geqslant 10$ в намеченном количестве минимальных выборок без перерыва. Продолжать такую процедуру до тех пор, пока либо не будет исчерпан намеченный объем выборки, либо (если есть необходимость и возможность) можно пренебречь намеченными границами и увеличивать объем выборки, пока не будет исчерпан перечень разделов.

Такой алгоритм может быть использован как для ручного составления ЧС, так и в качестве основы для разработки соответствующего машинного варианта. Думается, что преимущество его перед традиционными способами организации выборки состоит в том, что он предусматривает планомерное выделение достоверной лексики из всех проанализированных разделов выборки и в то же время оставляет возможность пополнения словаря при условии, что составитель сообщает все необходимые для этого сведения:

а) перечень обследованных областей семантического пространства;

б) размер минимальной выборки;

в) регистрируемые единицы;

г) среднее количество единиц в минимальной выборке;

д) значение $k^*_{крит}$;

е) принятое количество текстов, на которых должно выполняться условие $k_i^* \geqslant k^*_{крит}$ без перерыва.

---

[6] Следует отметить, что значение $k^*_{крит}$ пока определено на основании данных очень небольшого количества частотных словарей и требует уточнения. Уточнения требует и то, какое количество членов последовательности $\{k_i^*\}_{i=2}^{n}$, устойчиво превышающих значение $k^*_{крит}$, считать достаточным для перехода к следующему разделу выборки.

СПИСОК УСЛОВНЫХ СОКРАЩЕНИЙ

ЛЕ - лексическая единица
МКИ - Международная классификация изобретений
с/у - словоупотребление
ЧС - частотный словарь
№ п/п - порядковый номер

ЛИТЕРАТУРА

I. Сутягина Л.М. Формирование выборочного корпуса текстов при составлении частотного словаря (качественная сторона формирования выборки)// Квантитативные методы отбора учебного материала по иностранному языку для неязыкового вуза. Свердловск, 1986.

2. Фрумкина Р.М. Статистические методы изучения лексики. М., 1964.

3. Алексеев П.М. Статистическая лексикография. Л., 1975.

4. Перебейнос В.И. Определение надежности данных частотного словаря // Квантитативная лингвистика и автоматический анализ текстов. Тарту, 1984.

5. Пиотровский Р.Г. Инженерная лингвистика и теория языка. Л., 1979.

6. Тулдава Ю.А. О теоретико-методологических основах квантитативно-системного анализа лексики (2): лингвистические аспекты исследования // Лингвистика текста и стилистика: Лингвистика XIУ. Тарту, 1981.

7. Алимов Ю.И. Элементы теории эксперимента: Измерение моментов случайных величин, векторов и процессов. Свердловск, 1976.

8. Сутягина Л.М. Оптимизация составления частотного словаря (проблема статистической однородности выборки): Автореф.дис.... канд.филол. наук. Л.,1985.

В.Н.Бычков
Ленинградский пединститут

О ЛИНГВОСТАТИСТИЧЕСКОЙ ОПТИМИЗАЦИИ ОБУЧЕНИЯ ЧТЕНИЮ НА ИНОСТРАННОМ ЯЗЫКЕ В УСЛОВИЯХ ПРЕЕМСТВЕННОСТИ МЕЖДУ ШКОЛОЙ И НЕЯЗЫКОВЫМ ВУЗОМ

Известно, что, для того чтобы рационально организовать обучение и самообучение иностранному языку, необходимо четко